On the Learning and Representation of Human Utility

Penglin Cai
Yuanpei College
Peking University
cpl@stu.pku.edu.cn

Abstract

Utility functions are vital to human decision-making process, representing a preference over inhomogeneous things and goods. However, it has always been a difficult open problem to utilize a unified computational framework of estimating, learning, and representing human utilities, since utility functions are internal and implicit to humans, and can be different across individuals. Inspired by multiple related studies, we aim to design possible ways of computational models to learn and represent human utility. In this essay, we also present the strengths and weaknesses of each method.

1 Introduction

Utility, which is first introduced in economics, has become a significant term and function in measuring human's preferences [9]. In economics, a rational consumer would always maximize their utility within limited budget constraints [7]. In addition to economics, we can also model utility as we model the possible world (*e.g.*, a Markov Process) [2, 5]. In this case, utility can be used to define the extent of human's satisfaction of the possible world. Suppose the utility function with respect to n variables and factors $x_1, x_2, ..., x_n$ is

$$f(x_1, x_2, ..., x_n) \tag{1}$$

and the limited constraint (cost) can be represented as

$$c(x_1, x_2, ..., x_n) = 0 \tag{2}$$

In this ideal case, we can utilize Lagrange Multiplier Method to find the analytical solution, through

$$\max f(x_1, x_2, ..., x_n) - \lambda c(x_1, x_2, ..., x_n) \tag{3}$$

where λ is the Lagrange multiplier.

However, in the real world we usually cannot represent utility in a certain function. On the one hand, this is due to the complexity of the real world. There are too many variables and factors to be found, and whether they are relevant to our objective or not remains unclear. On the other hand, utility can be implicit and internal to humans, resulting in the difficulty in obtaining results through observations. From this perspective, how to estimate, learn, and represent human utilities remains a difficult problem. Inspired by multiple related studies, we aim to propose possible ways to design feasible computational frameworks to learn and represent human utility.

2 Related Work

Naïve utility calculus. Jara *et al*. [6] proposes that the naïve utility calculus consists of a theory or a generative model which is embedded in a Bayesian framework, and supports predictions about future behaviors (setting the costs and rewards and deriving the resulting actions) and inferences about the causes of observed behaviors. However, reasoning about decision-making in the real world has several complications that the idealized naïve utility calculus cannot handle. These complications

reveal more sophisticated aspects or the naïve utility calculus that give it traction and point to ways in which commonsense psychology may develop.

Reinforcement learning based on preferences and examples. Preference-based Reinforcement Learning (PbRL) [13] replaces reward values in traditional reinforcement learning by preferences to better elicit human opinion on the target objective, especially when numerical reward values are hard to design or interpret. Under the setting that preferences are stochastic, and the preference probability relates to the hidden reward values, the proposed PbRL algorithms are able to identify the best policy up to a high accuracy with high probability.

Inspired by the idea of devising RL algorithms that enable users to specify tasks simply by simply providing examples of successful outcomes, recursive classification of examples (RCE) [4] was proposed aimed at maximizing the future probability of the successful outcome examples. Different from other similar work, RCE directly learned a value function from transitions and successful outcomes, without learning the intermediate reward function. Experiments showed that RCE outperformed prior methods that learn explicit reward functions.

Sadigh *et al*. [10] built on work in label ranking and proposed to learn from preferences and/or comparisons - the person provided the system with a relative preference between two trajectories. Therefore, the learned reward function strongly depends on what environments and trajectories were experienced during the training phase.

Reinforcement learning with reward to represent utility. Silver *et al*. [11] hypothesised that intelligence and its associated abilities can be understood as subserving the maximization of reward, and reward is enough to drive a behaviour that exhibits abilities studied in natural and artificial intelligence (Figure 1).

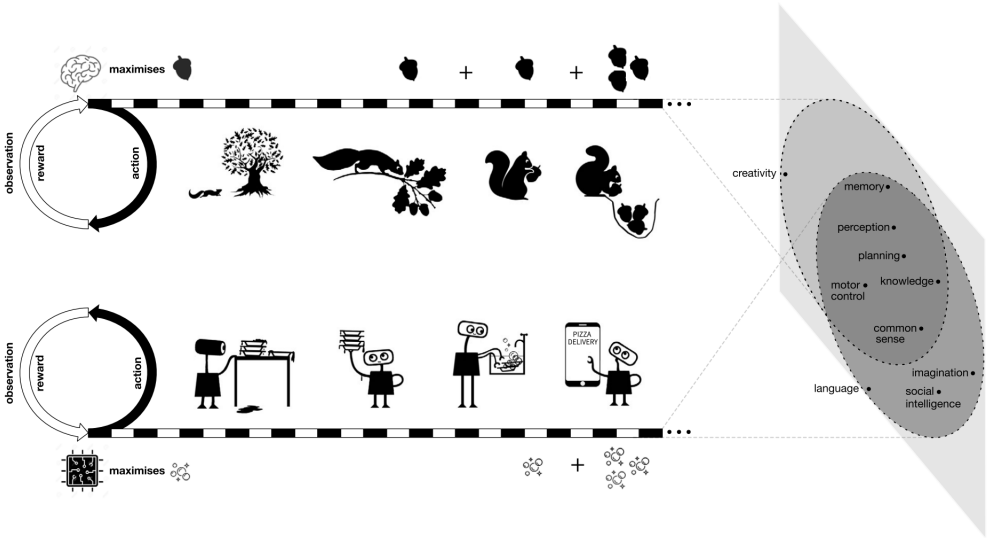

Figure 1: The reward-is-enough hypothesis postulates that intelligence, and its associated abilities, can be understood as subserving the maximisation of reward by an agent acting in its environment [11]. For example, a squirrel acts so as to maximise its consumption of food (top, reward depicted by acorn symbol), or a kitchen robot acts to maximise cleanliness (bottom, reward depicted by bubble symbol). To achieve these goals, complex behaviours are required that exhibit a wide variety of abilities associated with intelligence (depicted on the right as a projection from an agent's stream of experience onto a set of abilities expressed within that experience).

Abel *et al*. [1] proved that while reward can express many tasks, there existed instances of each task type that no Markov reward function can capture. Then a set of polynomial-time algorithms were provided to construct a Markov reward function that allows an agent to optimize tasks of three types -

- a set of acceptable behaviors;
- a partial ordering over behaviors;
- a partial ordering over trajectories.

with an empirical study for corroboration and illustration under a proposed framework (Figure 2).

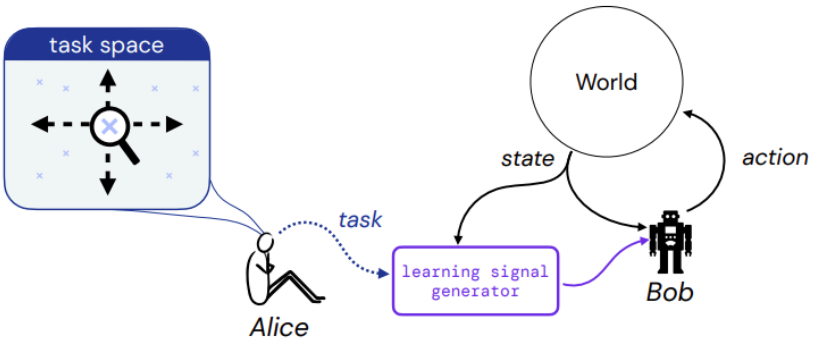

Figure 2: The framework. Alice, Bob, and the artifacts of task definition (blue) and task expression (purple).

Christiano *et al.* [3] explored goals defined in terms of human preferences between pairs of trajectory segments, and showed that such approach can effectively solve complex RL tasks without access to the reward function. Experimental results in the environments including Atari games and simulated robot locomotion showed that complex novel behaviors can be successfully trained within a flexible amount of time.

3 Possible Computational Methods

Inspired by previous work, we aim to propose a potential framework of computational models that can estimate, learn, and represent human utility. In the meanwhile, we will delve into each method about their advantages and disadvantages in data collection, generalization, and efficiency.

3.1 Utilizing the Partial Order Relationship

Sometimes the subjects within the scope of our discussion naturally have a relationship of partial order. For instance, prices of homogeneous economic goods can influence human's preference and utility [8]. With respect to a fixed quality of goods, lower prices stands for higher utility. Another example is the principle of "the more, the better". For homogeneous goods with a fixed price, having more of this kind of goods indicates higher utility [12].

In this aspect, the next step is to focus on learning a possible relationship of partial order, which also indicates a shortcoming of this method - partial order relationship may only exist among things that belong to the same category. Totally unrelated things (*e.g.*, apples and elephants), can hardly have partial order relationships inbetween.

In terms of learning a relationship of partial order, this can mainly rely on common sense, both physically and socially. From a generalized aspect, all measurement metrics and values have their corresponding partial order relationships. Supervised learning from a limited dataset in a specific domain can also be helpful with respect to this domain, be may become poor in generalizability.

3.2 Learning Preference from Observations and Inferences

Suppose a child is sitting in front of two toys, left and right. When you give the left one to this child, he/she begins to cry; on getting the right one, he/she laughs. This observation may indicate that this child prefers the right toy to the left one.

From a more generalized setting, given the set of all available (affordable) choices $A = \{x_1, x_2, ...x_n\}$ of a person, if he/she chooses $x_i \in A$, then it is indicated that he/she prefers x_i to others. Therefore, we can build a model to learn from numerous observations of human choices. Then through a simple inference, the preferences of the population can be indicated through a statistical calculation.

This method can be relatively accurate, since statistical patterns are always representative. However, constructing such a huge dataset may consume too much labor, effort, and time.

3.3 Learning from Statistics in Existing Resources

Existing resources, such as corpus used in natural language processing, contains numerous information of human preference and utility. Consider this example:

> Life is dear , love is dearer . Both can be given up for freedom. – Petofi Sandor

From this quote, we can know that from the perspective of Petofi Sandor, the utility of having freedom is much higher than possessing either of the former two.

Following this method, we can design a model to capture relevant information from large quantities of corpuses. This method can save much labor and time, since all the dataset already exist. However, one drawback is that the corpus can be polluted, and needs cleaning and post-processing.

4 Conclusion

In this essay, we reviewed previous work and research on learning and representing human utility via different methods, including naïve utility calculus, RL based on examples of preferences, and RL with reward representations. The implicitness and internality of utility function make it difficult to estimate and represent human utility. However, we manage to propose several possible ways of constructing computational methods to model the human utility. We also discuss the advantages and disadvantages of each method, hoping to be of inspiration for further research.

References

[1] David Abel, Will Dabney, Anna Harutyunyan, Mark K Ho, Michael Littman, Doina Precup, and Satinder Singh. On the expressivity of markov reward. *Advances in Neural Information Processing Systems*, 34:7799–7812, 2021. 2

[2] Wallace E Armstrong. The determinateness of the utility function. *The Economic Journal*, 49 (195):453–467, 1939. 1

[3] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017. 3

[4] Ben Eysenbach, Sergey Levine, and Russ R Salakhutdinov. Replacing rewards with examples: Example-based policy search via recursive classification. *Advances in Neural Information Processing Systems*, 34:11541–11552, 2021. 2

[5] Hendrik S Houthakker. Revealed preference and the utility function. *Economica*, 17(66): 159–174, 1950. 1

[6] Julian Jara-Ettinger, Hyowon Gweon, Laura E Schulz, and Joshua B Tenenbaum. The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences*, 20(8):589–604, 2016. 1

[7] Boris Mirkin. Reinterpreting the category utility function. *Machine Learning*, 45:219–228, 2001. 1

[8] James M Poterba and Julio J Rotemberg. Money in the utility function: An empirical implementation, 1986. 3

[9] Trout Rader. The existence of a utility function to represent preferences. *The Review of Economic Studies*, 30(3):229–232, 1963. 1

[10] Dorsa Sadigh, Anca D Dragan, Shankar Sastry, and Sanjit A Seshia. *Active preference-based learning of reward functions*. 2017. 2

[11] David Silver, Satinder Singh, Doina Precup, and Richard S Sutton. Reward is enough. *Artificial Intelligence*, 299:103535, 2021. 2

[12] AA Vasin, PA Vasina, and T Yu Ruleva. On organization of markets of homogeneous goods. *Journal of Computer and Systems Sciences International*, 46:93–106, 2007. 3

[13] Yichong Xu, Ruosong Wang, Lin Yang, Aarti Singh, and Artur Dubrawski. Preference-based reinforcement learning with finite-time guarantees. *Advances in Neural Information Processing Systems*, 33:18784–18794, 2020. 2